06: Correlation and Intro to Regression

Stat 120 | Fall 2025

Prof Amanda Luby

This dataset gives education-related data for the 50 states and the District of Columbia. The variables are:

- `region`: West, Northeast, Midwest, South
- `pop`: Population, in 1,000's
- `verbal` and `math`: average SAT verbal and math scores
- `taken`: percent of students taking the SAT
- `noHS`: percent of population with no high school diploma
- `teachersPay`: median teacher salary, in 1,000's

0. Load code libraries and the data and make sure you can view it. What is each case?

```r
library(tidyverse)
sat <- read.csv("http://math.carleton.edu/Stat120/RLabManual/sat.csv")
```

We're going to investigate the relationship between math SAT scores (`math`) and the percentage of high school students who took the SAT.

1. *Before looking at the data*, do expect there to be a positive, negative, or *no* relationship? Why?

2. Make a scatterplot of `math` on the y-axis and `taken` on the x-axis with the line of best fit included (see notes from today for the line of code to include). What do you notice?

3. Use the `lm()` command to find the equation for this line. *Be careful about the X and Y variables!* Interpret the slope and intercept in context.

4. Find the correlation for this relationship.

5. Color the scatterplot by `region`. What do you notice?

The code chunk below creates a new dataset called `sat_northeast` which filters to only the Northeast states (`==` is code for "equals").

```
sat_northeast <- sat %>%
  filter(region == "Northeast")
```

6. Make a scatterplot of `math` on the y-axis and `taken` on the x-axis with the line of best fit included.

7. Use the `lm()` command to find the equation for this line. How does it compare to your line from (3)?

8. Find the correlation for this relationship. How does it compare to the correlation for the whole dataset?

Note: this data is adapted from Ch 3.4 of the Lab Manual. You can find most of the code solutions there if your group gets stuck!